

# Linear Regression

---

Nipun Batra and the teaching staff

July 21, 2025

IIT Gandhinagar

## Setup

---

# Linear Regression

- Output is continuous in nature.

# Linear Regression

- Output is continuous in nature.
- Examples of linear systems:

# Linear Regression

- Output is continuous in nature.
- Examples of linear systems:
  - $F = ma$

# Linear Regression

- Output is continuous in nature.
- Examples of linear systems:
  - $F = ma$
  - $v = u + at$

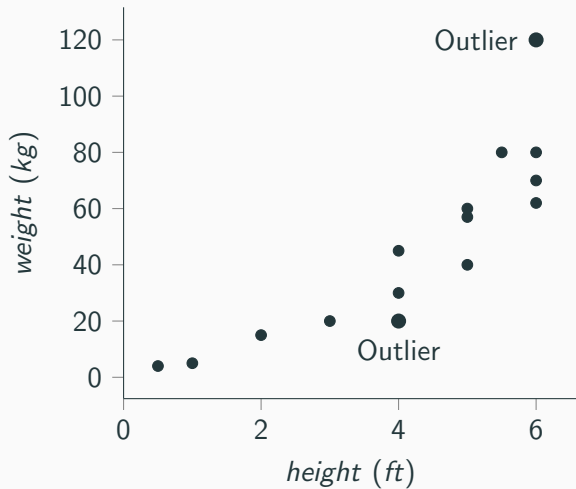
## Task at hand

- TASK: Predict  $\text{Weight} = f(\text{height})$

Height	Weight
3	29
4	35
5	39
2	20
6	41
7	?
8	?
1	?

The first part of the dataset is the training points. The latter ones are testing points.

# Scatter Plot





## Matrix representation of the expression

- $weight_1 \approx \theta_0 + \theta_1 \cdot height_1$

# Matrix representation of the expression

- $weight_1 \approx \theta_0 + \theta_1 \cdot height_1$
- $weight_2 \approx \theta_0 + \theta_1 \cdot height_2$

# Matrix representation of the expression

- $weight_1 \approx \theta_0 + \theta_1 \cdot height_1$
- $weight_2 \approx \theta_0 + \theta_1 \cdot height_2$
- $weight_N \approx \theta_0 + \theta_1 \cdot height_N$

# Matrix representation of the expression

- $weight_1 \approx \theta_0 + \theta_1 \cdot height_1$
- $weight_2 \approx \theta_0 + \theta_1 \cdot height_2$
- $weight_N \approx \theta_0 + \theta_1 \cdot height_N$

# Matrix representation of the expression

- $weight_1 \approx \theta_0 + \theta_1 \cdot height_1$
- $weight_2 \approx \theta_0 + \theta_1 \cdot height_2$
- $weight_N \approx \theta_0 + \theta_1 \cdot height_N$

$$weight_i \approx \theta_0 + \theta_1 \cdot height_i$$

## Matrix representation of the expression

$$\begin{bmatrix} weight_1 \\ weight_2 \\ \dots \\ weight_N \end{bmatrix} = \begin{bmatrix} 1 & height_1 \\ 1 & height_2 \\ \dots & \dots \\ 1 & height_N \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

## Matrix representation of the expression

$$\begin{bmatrix} weight_1 \\ weight_2 \\ \dots \\ weight_N \end{bmatrix} = \begin{bmatrix} 1 & height_1 \\ 1 & height_2 \\ \dots & \dots \\ 1 & height_N \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

$$\hat{\mathbf{y}}_{n \times 1} = \mathbf{X}_{n \times d} \boldsymbol{\theta}_{d \times 1}$$

## Matrix representation of the expression

$$\begin{bmatrix} weight_1 \\ weight_2 \\ \dots \\ weight_N \end{bmatrix} = \begin{bmatrix} 1 & height_1 \\ 1 & height_2 \\ \dots & \dots \\ 1 & height_N \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

$$\hat{\mathbf{y}}_{n \times 1} = \mathbf{X}_{n \times d} \boldsymbol{\theta}_{d \times 1}$$

- $\theta_0$  - Bias Term/Intercept Term



## Matrix representation of the expression

$$\begin{bmatrix} weight_1 \\ weight_2 \\ \dots \\ weight_N \end{bmatrix} = \begin{bmatrix} 1 & height_1 \\ 1 & height_2 \\ \dots & \dots \\ 1 & height_N \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

$$\hat{\mathbf{y}}_{n \times 1} = \mathbf{X}_{n \times d} \boldsymbol{\theta}_{d \times 1}$$

- $\theta_0$  - Bias Term/Intercept Term
- $\theta_1$  - Slope

## Extension to multiple dimensions

In the previous example  $y = f(x)$ , where  $x$  is one-dimensional.

## Extension to multiple dimensions

In the previous example  $y = f(x)$ , where  $x$  is one-dimensional.  
Examples in multiple dimensions.

## Extension to multiple dimensions

In the previous example  $y = f(x)$ , where  $x$  is one-dimensional.

Examples in multiple dimensions.

One example is to predict the water demand of the IITGN campus

## Extension to multiple dimensions

In the previous example  $y = f(x)$ , where  $x$  is one-dimensional.

Examples in multiple dimensions.

One example is to predict the water demand of the IITGN campus

$$\text{Demand} = f(\# \text{ occupants, Temperature})$$

## Extension to multiple dimensions

In the previous example  $y = f(x)$ , where  $x$  is one-dimensional.

Examples in multiple dimensions.

One example is to predict the water demand of the IITGN campus

$$\text{Demand} = f(\# \text{ occupants, Temperature})$$

$$\text{Demand} = \text{Base Demand} + K_1 * \# \text{ occupants} + K_2 * \text{Temperature}$$

We hope to:

- Learn  $f$ :  $Demand = f(\#occupants, Temperature)$

We hope to:

- Learn  $f$ :  $Demand = f(\#occupants, Temperature)$
- From training dataset



We hope to:

- Learn  $f$ :  $Demand = f(\#occupants, Temperature)$
- From training dataset
- To predict the condition for the testing set

# Linear Relationship

We have

- $x_i = \begin{bmatrix} \textit{Temperature}_i \\ \textit{\#Occupants}_i \end{bmatrix}$

# Linear Relationship

We have

- $x_i = \begin{bmatrix} \text{Temperature}_i \\ \# \text{Occupants}_i \end{bmatrix}$
- Estimated demand for  $i^{\text{th}}$  sample is  
 $\hat{\text{demand}}_i = \theta_0 + \theta_1 \text{Temperature}_i + \theta_2 \text{Occupants}_i$

# Linear Relationship

We have

- $x_i = \begin{bmatrix} \text{Temperature}_i \\ \# \text{Occupants}_i \end{bmatrix}$
- Estimated demand for  $i^{\text{th}}$  sample is
$$\hat{\text{demand}}_i = \theta_0 + \theta_1 \text{Temperature}_i + \theta_2 \text{Occupants}_i$$
- $\hat{\text{demand}}_i = x_i'^T \theta$

# Linear Relationship

We have

- $x_i = \begin{bmatrix} \text{Temperature}_i \\ \# \text{Occupants}_i \end{bmatrix}$
- Estimated demand for  $i^{\text{th}}$  sample is
$$\hat{\text{demand}}_i = \theta_0 + \theta_1 \text{Temperature}_i + \theta_2 \text{Occupants}_i$$
- $\hat{\text{demand}}_i = x_i'^T \theta$
- where  $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$

# Linear Relationship

We have

- $x_i = \begin{bmatrix} \text{Temperature}_i \\ \# \text{Occupants}_i \end{bmatrix}$
- Estimated demand for  $i^{\text{th}}$  sample is
$$\hat{\text{demand}}_i = \theta_0 + \theta_1 \text{Temperature}_i + \theta_2 \text{Occupants}_i$$
- $\hat{\text{demand}}_i = x_i'^T \theta$
- where  $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$
- and  $x_i' = \begin{bmatrix} 1 \\ \text{Temperature}_i \\ \# \text{Occupants}_i \end{bmatrix} = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$

# Linear Relationship

We have

- $x_i = \begin{bmatrix} \text{Temperature}_i \\ \text{\#Occupants}_i \end{bmatrix}$
- Estimated demand for  $i^{\text{th}}$  sample is
$$\hat{\text{demand}}_i = \theta_0 + \theta_1 \text{Temperature}_i + \theta_2 \text{Occupants}_i$$
- $\hat{\text{demand}}_i = x_i'^T \theta$
- where  $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$
- and  $x_i' = \begin{bmatrix} 1 \\ \text{Temperature}_i \\ \text{\#Occupants}_i \end{bmatrix} = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$
- Notice the transpose in the equation! This is because  $x_i$  is a column vector

## We can expect the following

- Demand increases, if # occupants increases, then  $\theta_2$  is likely to be positive



## We can expect the following

- Demand increases, if # occupants increases, then  $\theta_2$  is likely to be positive
- Demand increases, if temperature increases, then  $\theta_1$  is likely to be positive

## We can expect the following

- Demand increases, if # occupants increases, then  $\theta_2$  is likely to be positive
- Demand increases, if temperature increases, then  $\theta_1$  is likely to be positive
- Base demand is independent of the temperature and the # occupants, but, likely positive, thus  $\theta_0$  is likely positive.

# Normal Equation

---

# Generalized Linear Regression Format

- Assuming  $N$  samples for training

# Generalized Linear Regression Format

- Assuming  $N$  samples for training
- # Features =  $M$

# Generalized Linear Regression Format

- Assuming  $N$  samples for training
- # Features =  $M$

# Generalized Linear Regression Format

- Assuming  $N$  samples for training
- # Features =  $M$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}_{N \times 1} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}_{N \times (M+1)} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}_{(M+1) \times 1}$$

# Generalized Linear Regression Format

- Assuming  $N$  samples for training
- # Features =  $M$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}_{N \times 1} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}_{N \times (M+1)} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}_{(M+1) \times 1}$$

$$\hat{Y} = X\theta$$



## Relationships between feature and target variables

- There could be different  $\theta_0, \theta_1 \dots \theta_M$ . Each of them can represents a relationship.

## Relationships between feature and target variables

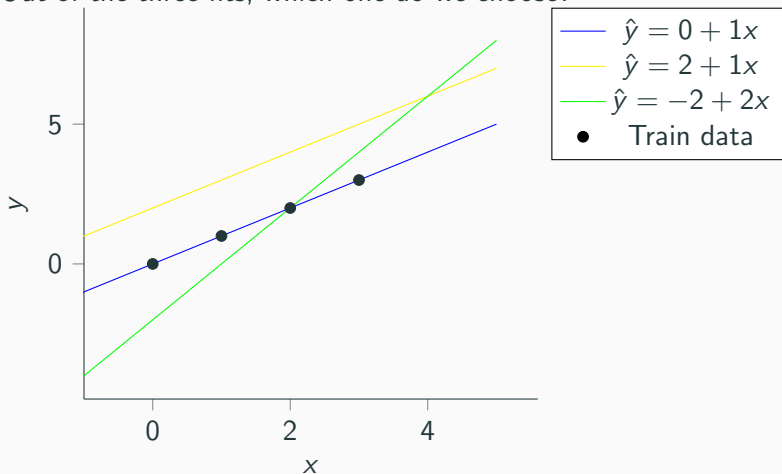
- There could be different  $\theta_0, \theta_1 \dots \theta_M$ . Each of them can represents a relationship.
- Given multiples values of  $\theta_0, \theta_1 \dots \theta_M$  how to choose which is the best?

## Relationships between feature and target variables

- There could be different  $\theta_0, \theta_1 \dots \theta_M$ . Each of them can represents a relationship.
- Given multiples values of  $\theta_0, \theta_1 \dots \theta_M$  how to choose which is the best?
- Let us consider an example in 2d

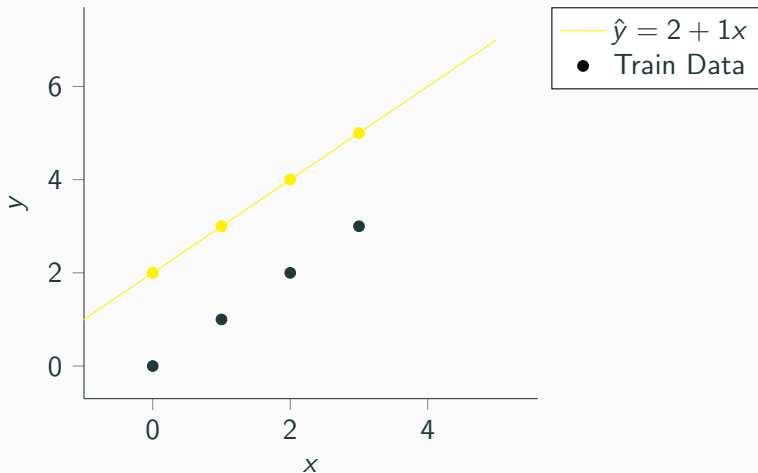
## Relationships between feature and target variables

Out of the three fits, which one do we choose?



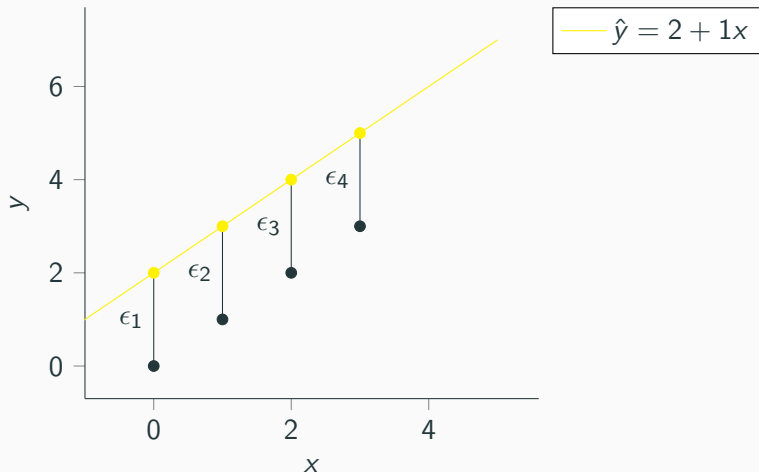
# Relationships between feature and target variables

We have  $\hat{y} = 2 + 1x$  as one relationship.



# Relationships between feature and target variables

How far is our estimated  $\hat{y}$  from ground truth  $y$ ?



- $y_i = \hat{y}_i + \epsilon_i$  where  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

## Error terms

- $y_i = \hat{y}_i + \epsilon_i$  where  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- $y_i$  denotes the ground truth for  $i^{th}$  sample



## Error terms

- $y_i = \hat{y}_i + \epsilon_i$  where  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- $y_i$  denotes the ground truth for  $i^{th}$  sample
- $\hat{y}_i$  denotes the prediction for  $i^{th}$  sample, where  $\hat{y}_i = \mathbf{x}_i^\top \boldsymbol{\theta}$

## Error terms

- $y_i = \hat{y}_i + \epsilon_i$  where  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- $y_i$  denotes the ground truth for  $i^{th}$  sample
- $\hat{y}_i$  denotes the prediction for  $i^{th}$  sample, where  $\hat{y}_i = \mathbf{x}_i^\top \boldsymbol{\theta}$
- $\epsilon_i$  denotes the error/residual for  $i^{th}$  sample

## Error terms

- $y_i = \hat{y}_i + \epsilon_i$  where  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- $y_i$  denotes the ground truth for  $i^{th}$  sample
- $\hat{y}_i$  denotes the prediction for  $i^{th}$  sample, where  $\hat{y}_i = \mathbf{x}_i^\top \boldsymbol{\theta}$
- $\epsilon_i$  denotes the error/residual for  $i^{th}$  sample
- $\theta_0, \theta_1$ : The parameters of the linear regression

## Error terms

- $y_i = \hat{y}_i + \epsilon_i$  where  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- $y_i$  denotes the ground truth for  $i^{th}$  sample
- $\hat{y}_i$  denotes the prediction for  $i^{th}$  sample, where  $\hat{y}_i = \mathbf{x}_i^\top \boldsymbol{\theta}$
- $\epsilon_i$  denotes the error/residual for  $i^{th}$  sample
- $\theta_0, \theta_1$ : The parameters of the linear regression
- $\epsilon_i = y_i - \hat{y}_i$

## Error terms

- $y_i = \hat{y}_i + \epsilon_i$  where  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- $y_i$  denotes the ground truth for  $i^{th}$  sample
- $\hat{y}_i$  denotes the prediction for  $i^{th}$  sample, where  $\hat{y}_i = \mathbf{x}_i^\top \boldsymbol{\theta}$
- $\epsilon_i$  denotes the error/residual for  $i^{th}$  sample
- $\theta_0, \theta_1$ : The parameters of the linear regression
- $\epsilon_i = y_i - \hat{y}_i$
- $\epsilon_i = y_i - (\theta_0 + x_i \cdot \theta_1)$

- $|\epsilon_1|, |\epsilon_2|, |\epsilon_3|, \dots$  should be small.

- $|\epsilon_1|, |\epsilon_2|, |\epsilon_3|, \dots$  should be small.
- minimize  $\epsilon_1^2 + \epsilon_2^2 + \dots + \epsilon_N^2$  -  $L_2$  Norm

- $|\epsilon_1|, |\epsilon_2|, |\epsilon_3|, \dots$  should be small.
- minimize  $\epsilon_1^2 + \epsilon_2^2 + \dots + \epsilon_N^2$  -  $L_2$  Norm
- minimize  $|\epsilon_1| + |\epsilon_2| + \dots + |\epsilon_n|$  -  $L_1$  Norm



# Normal Equation

## Normal Equation

$$Y = X\theta + \epsilon$$

## Normal Equation

$$Y = X\theta + \epsilon$$

To Learn:  $\theta$

## Normal Equation

$$Y = X\theta + \epsilon$$

To Learn:  $\theta$

Objective: minimize  $\epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_N^2$

# Normal Equation

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

# Normal Equation

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

Objective: Minimize  $\epsilon^T \epsilon$

## Derivation of Normal Equation

$$\epsilon = \mathbf{y} - \mathbf{X}\theta$$

$$\epsilon^\top \epsilon = (\mathbf{y} - \mathbf{X}\theta)^\top (\mathbf{y} - \mathbf{X}\theta)$$

$$= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X}\theta + \theta^\top \mathbf{X}^\top \mathbf{X}\theta$$

This is what we wish to minimize

## Minimizing the objective function

$$\frac{\partial \epsilon^\top \epsilon}{\partial \theta} = 0$$

- $\frac{\partial}{\partial \theta} \mathbf{y}^\top \mathbf{y} = 0$

Substitute the values in the top equation



## Minimizing the objective function

$$\frac{\partial \epsilon^\top \epsilon}{\partial \theta} = 0$$

- $\frac{\partial}{\partial \theta} \mathbf{y}^\top \mathbf{y} = 0$
- $\frac{\partial}{\partial \theta} (-2\mathbf{y}^\top \mathbf{X}\theta) = -2\mathbf{X}^\top \mathbf{y}$

Substitute the values in the top equation

## Minimizing the objective function

$$\frac{\partial \epsilon^\top \epsilon}{\partial \theta} = 0$$

- $\frac{\partial}{\partial \theta} \mathbf{y}^\top \mathbf{y} = 0$
- $\frac{\partial}{\partial \theta} (-2\mathbf{y}^\top \mathbf{X}\theta) = -2\mathbf{X}^\top \mathbf{y}$
- $\frac{\partial}{\partial \theta} (\theta^\top \mathbf{X}^\top \mathbf{X}\theta) = 2\mathbf{X}^\top \mathbf{X}\theta$

Substitute the values in the top equation

## Normal Equation derivation

$$0 = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X} \boldsymbol{\theta}$$

$$\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X} \boldsymbol{\theta}$$

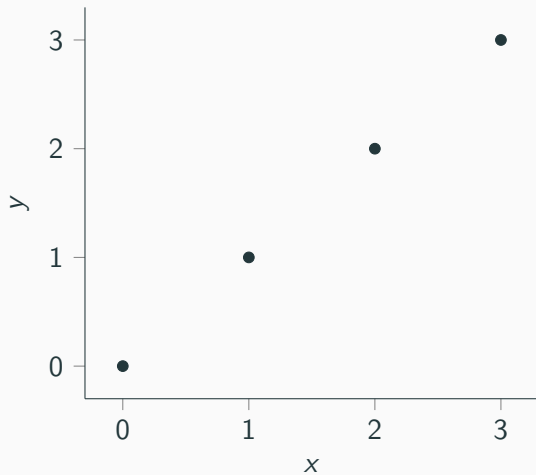
$$\hat{\boldsymbol{\theta}}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

## Worked out example

x	y
0	0
1	1
2	2
3	3

Given the data above, find  $\theta_0$  and  $\theta_1$ .

# Scatter Plot



## Worked out example

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

$$\mathbf{X}^{\top} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix}$$

$$\mathbf{X}^{\top} \mathbf{X} = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix}$$

Given the data above, find  $\theta_0$  and  $\theta_1$ .

## Worked out example

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{20} \begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix}$$

$$\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

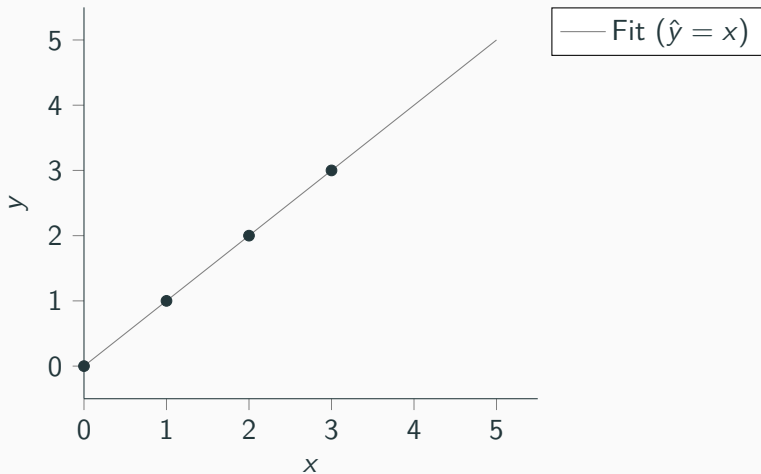
## Worked out example

$$\theta = (X^T X)^{-1} (X^T y)$$

$$\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \frac{1}{20} \begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix} \begin{bmatrix} 6 \\ 14 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$



# Scatter Plot

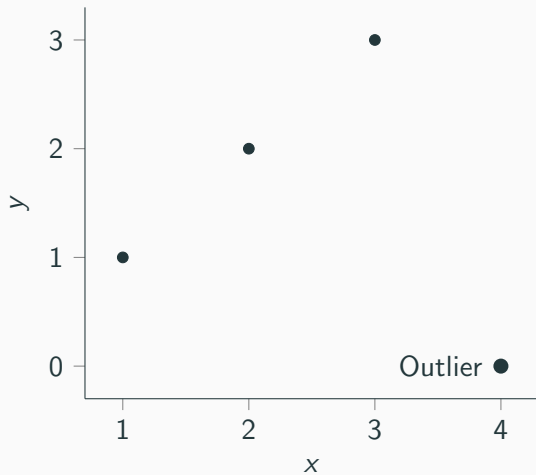


## Effect of outlier

x	y
1	1
2	2
3	3
4	0

Compute the  $\theta_0$  and  $\theta_1$ .

# Scatter Plot



## Worked out example

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

Given the data above, find  $\theta_0$  and  $\theta_1$ .

## Worked out example

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix}$$

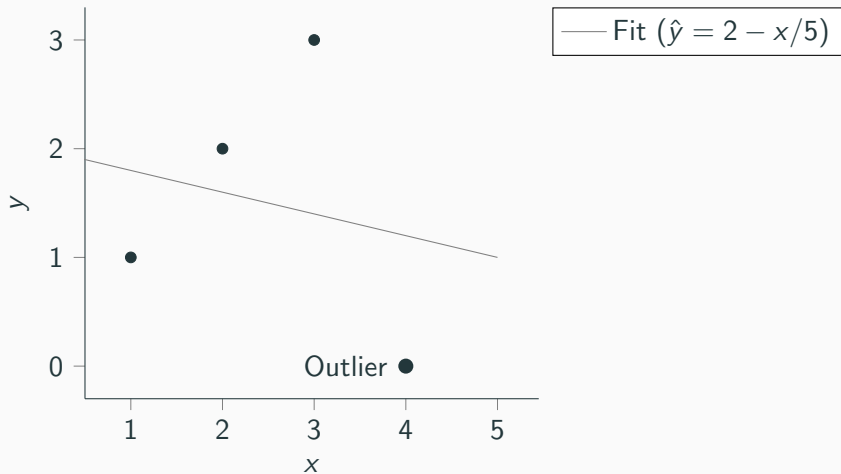
$$\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

## Worked out example

$$\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^\top \mathbf{y})$$

$$\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 2 \\ (-1/5) \end{bmatrix}$$

# Scatter Plot



# Basis Expansion

---



# Variable Transformation

Transform the data, by including the higher power terms in the feature space.

t	s
0	0
1	6
3	24
4	36

The above table represents the data before transformation

## Variable Transformation

Add the higher degree features to the previous table

t	$t^2$	s
0	0	0
1	1	6
3	9	24
4	16	36

## Variable Transformation

Add the higher degree features to the previous table

t	$t^2$	s
0	0	0
1	1	6
3	9	24
4	16	36

The above table represents the data after transformation

## Variable Transformation

Add the higher degree features to the previous table

t	$t^2$	s
0	0	0
1	1	6
3	9	24
4	16	36

The above table represents the data after transformation

Now, we can write  $\hat{s} = f(t, t^2)$

## Variable Transformation

Add the higher degree features to the previous table

t	$t^2$	s
0	0	0
1	1	6
3	9	24
4	16	36

The above table represents the data after transformation

Now, we can write  $\hat{s} = f(t, t^2)$

Other transformations:  $\log(x)$ ,  $x_1 \times x_2$

## A big caveat: Linear in what?!<sup>1</sup>

1.  $\hat{s} = \theta_0 + \theta_1 * t$  is linear

---

<sup>1</sup><https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression>

## A big caveat: Linear in what?!<sup>1</sup>

1.  $\hat{s} = \theta_0 + \theta_1 * t$  is linear
2. Is  $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$  linear?

---

<sup>1</sup><https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression>

## A big caveat: Linear in what?!<sup>1</sup>

1.  $\hat{s} = \theta_0 + \theta_1 * t$  is linear
2. Is  $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$  linear?
3. Is  $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$  linear?

---

<sup>1</sup><https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression>



## A big caveat: Linear in what?!<sup>1</sup>

1.  $\hat{s} = \theta_0 + \theta_1 * t$  is linear
2. Is  $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$  linear?
3. Is  $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$  linear?
4. Is  $\hat{s} = \theta_0 + \theta_1 * t + e^{\theta_2} * t$  linear?

---

<sup>1</sup><https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression>

## A big caveat: Linear in what?!<sup>1</sup>

1.  $\hat{s} = \theta_0 + \theta_1 * t$  is linear
2. Is  $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$  linear?
3. Is  $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$  linear?
4. Is  $\hat{s} = \theta_0 + \theta_1 * t + e^{\theta_2} * t$  linear?
5. All except #4 are linear models!

---

<sup>1</sup><https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression>

## A big caveat: Linear in what?!<sup>1</sup>

1.  $\hat{s} = \theta_0 + \theta_1 * t$  is linear
2. Is  $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$  linear?
3. Is  $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$  linear?
4. Is  $\hat{s} = \theta_0 + \theta_1 * t + e^{\theta_2} * t$  linear?
5. All except #4 are linear models!
6. Linear refers to the relationship between the parameters that you are estimating ( $\theta$ ) and the outcome

---

<sup>1</sup><https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression>

- Linear regression only refers to linear in the parameters

- Linear regression only refers to linear in the parameters
- We can perform an arbitrary nonlinear transformation  $\phi(x)$  of the inputs  $x$  and then linearly combine the components of this transformation.

# Basis Functions

- Linear regression only refers to linear in the parameters
- We can perform an arbitrary nonlinear transformation  $\phi(x)$  of the inputs  $x$  and then linearly combine the components of this transformation.
- $\phi : \mathbb{R}^D \rightarrow \mathbb{R}^K$  is called the basis function

Some examples of basis functions:

- Polynomial basis:  $\phi(x) = \{1, x, x^2, x^3, \dots\}$

Some examples of basis functions:

- Polynomial basis:  $\phi(x) = \{1, x, x^2, x^3, \dots\}$
- Fourier basis:  $\phi(x) = \{1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots\}$



Some examples of basis functions:

- Polynomial basis:  $\phi(x) = \{1, x, x^2, x^3, \dots\}$
- Fourier basis:  $\phi(x) = \{1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots\}$
- Gaussian basis:  $\phi(x) = \{1, \exp(-\frac{(x-\mu_1)^2}{2\sigma^2}), \exp(-\frac{(x-\mu_2)^2}{2\sigma^2}), \dots\}$

Some examples of basis functions:

- Polynomial basis:  $\phi(x) = \{1, x, x^2, x^3, \dots\}$
- Fourier basis:  $\phi(x) = \{1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots\}$
- Gaussian basis:  $\phi(x) = \{1, \exp(-\frac{(x-\mu_1)^2}{2\sigma^2}), \exp(-\frac{(x-\mu_2)^2}{2\sigma^2}), \dots\}$
- Sigmoid basis:  $\phi(x) = \{1, \sigma(x - \mu_1), \sigma(x - \mu_2), \dots\}$  where  $\sigma(x) = \frac{1}{1+e^{-x}}$

# Geometric Interpretation

---

# Linear Combination of Vectors

Let  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_i$  be vectors in  $\mathbb{R}^D$ , where  $D$  denotes the dimensions.

# Linear Combination of Vectors

Let  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_i$  be vectors in  $\mathbb{R}^D$ , where  $D$  denotes the dimensions.

A linear combination of  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_i$  is of the following form

# Linear Combination of Vectors

Let  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_i$  be vectors in  $\mathbb{R}^D$ , where  $D$  denotes the dimensions.

A linear combination of  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_i$  is of the following form

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \alpha_3 \mathbf{v}_3 + \dots + \alpha_i \mathbf{v}_i$$

where  $\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_i \in \mathbb{R}$

## Span of vectors

Let  $v_1, v_2, \dots, v_i$  be vectors in  $\mathbb{R}^D$ , with  $D$  dimensions.

## Span of vectors

Let  $v_1, v_2, \dots, v_i$  be vectors in  $\mathbb{R}^D$ , with  $D$  dimensions.

The span of  $v_1, v_2, \dots, v_i$  is denoted by  $\text{SPAN}\{v_1, v_2, \dots, v_i\}$



## Span of vectors

Let  $v_1, v_2, \dots, v_i$  be vectors in  $\mathbb{R}^D$ , with  $D$  dimensions.

The span of  $v_1, v_2, \dots, v_i$  is denoted by  $\text{SPAN}\{v_1, v_2, \dots, v_i\}$

$$\{\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_i v_i \mid \alpha_1, \alpha_2, \dots, \alpha_i \in \mathbb{R}\}$$

## Span of vectors

Let  $v_1, v_2, \dots, v_i$  be vectors in  $\mathbb{R}^D$ , with  $D$  dimensions.

The span of  $v_1, v_2, \dots, v_i$  is denoted by  $\text{SPAN}\{v_1, v_2, \dots, v_i\}$

$$\{\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_i v_i \mid \alpha_1, \alpha_2, \dots, \alpha_i \in \mathbb{R}\}$$

It is the set of all vectors that can be generated by linear combinations of  $v_1, v_2, \dots, v_i$ .

## Span of vectors

Let  $v_1, v_2, \dots, v_i$  be vectors in  $\mathbb{R}^D$ , with  $D$  dimensions.

The span of  $v_1, v_2, \dots, v_i$  is denoted by  $\text{SPAN}\{v_1, v_2, \dots, v_i\}$

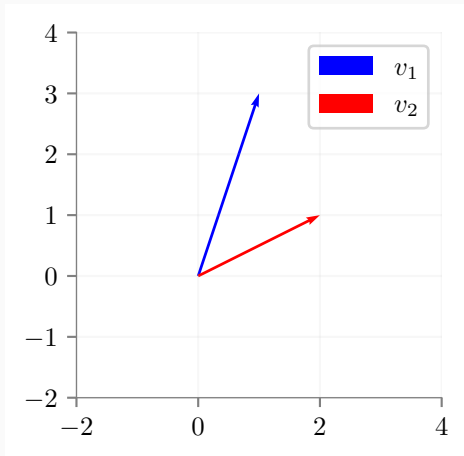
$$\{\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_i v_i \mid \alpha_1, \alpha_2, \dots, \alpha_i \in \mathbb{R}\}$$

It is the set of all vectors that can be generated by linear combinations of  $v_1, v_2, \dots, v_i$ .

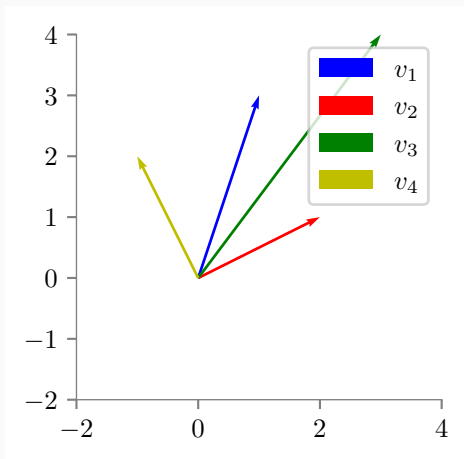
If we stack the vectors  $v_1, v_2, \dots, v_i$  as columns of a matrix  $V$ , then the span of  $v_1, v_2, \dots, v_i$  is given as  $V\alpha$  where  $\alpha \in \mathbb{R}^i$

## Example

Find the span of  $\left( \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix} \right)$



## Example

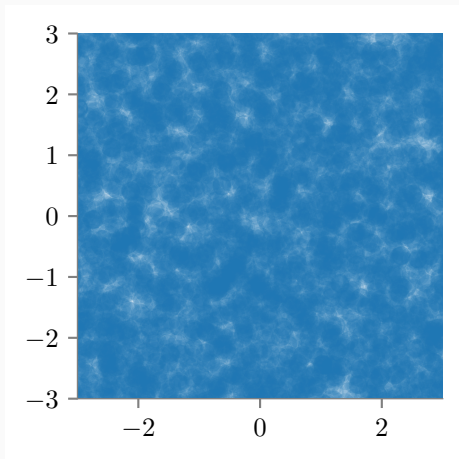


We have  $v_3 = v_1 + v_2$

We have  $v_4 = v_1 - v_2$

## Example

Simulating the above example in python using different values of  $\alpha_1$  and  $\alpha_2$



$$\text{Span}((v_1, v_2)) \in \mathcal{R}^2$$

## Example

Find the span of  $\left( \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right)$

## Example

Find the span of  $\left( \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right)$

Can we obtain a point  $(x, y)$  s.t.  $x = 3y$ ?



## Example

Find the span of  $\left( \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right)$

Can we obtain a point  $(x, y)$  s.t.  $x = 3y$ ?

No

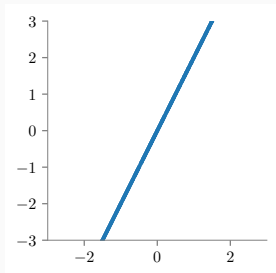
## Example

Find the span of  $\left( \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right)$

Can we obtain a point  $(x, y)$  s.t.  $x = 3y$ ?

No

Span of the above set is along the line  $y = 2x$

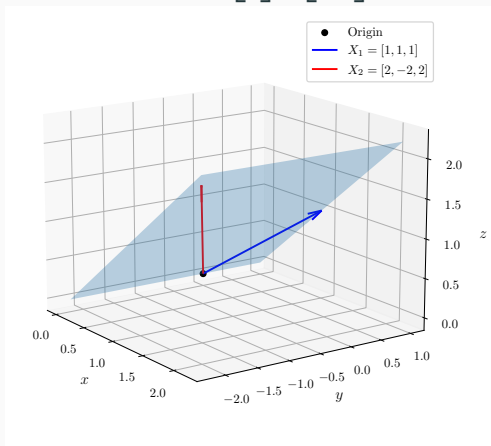


## Example

Find the span of  $\left( \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix} \right)$

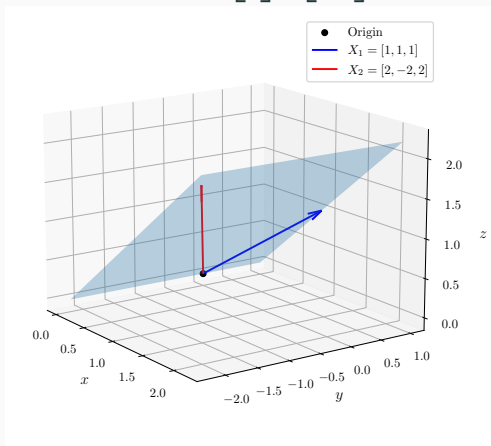
## Example

Find the span of  $\left( \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix} \right)$



## Example

Find the span of  $\left( \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix} \right)$



The span is the plane  $z = x$  or  $x_3 = x_1$

## Geometric Interpretation

Consider  $\mathbf{X}$  and  $\mathbf{y}$  as follows.

$$\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 8.8957 \\ 0.6130 \\ 1.7761 \end{pmatrix}$$

- We are trying to learn  $\boldsymbol{\theta}$  for  $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$  such that  $\|\mathbf{y} - \hat{\mathbf{y}}\|_2$  is minimised

## Geometric Interpretation

Consider  $\mathbf{X}$  and  $\mathbf{y}$  as follows.

$$\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 8.8957 \\ 0.6130 \\ 1.7761 \end{pmatrix}$$

- We are trying to learn  $\boldsymbol{\theta}$  for  $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$  such that  $\|\mathbf{y} - \hat{\mathbf{y}}\|_2$  is minimised
- Consider the two columns of  $\mathbf{X}$ . Can we write  $\mathbf{X}\boldsymbol{\theta}$  as the span of  $\left( \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix} \right)$ ?

# Geometric Interpretation

Consider  $\mathbf{X}$  and  $\mathbf{y}$  as follows.

$$\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 8.8957 \\ 0.6130 \\ 1.7761 \end{pmatrix}$$

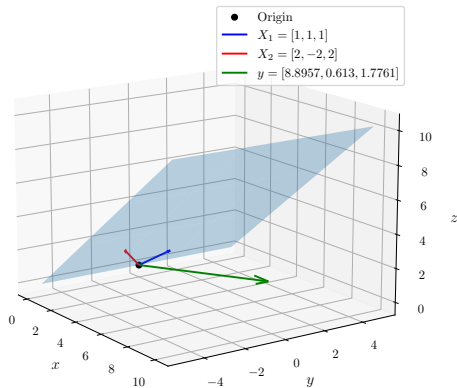
- We are trying to learn  $\boldsymbol{\theta}$  for  $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$  such that  $\|\mathbf{y} - \hat{\mathbf{y}}\|_2$  is minimised
- Consider the two columns of  $\mathbf{X}$ . Can we write  $\mathbf{X}\boldsymbol{\theta}$  as the span of  $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ ,  $\begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix}$ ?
- We wish to find  $\hat{\mathbf{y}}$  such that

$$\arg \min_{\hat{\mathbf{y}} \in \text{SPAN}\{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_D\}} \|\mathbf{y} - \hat{\mathbf{y}}\|_2$$

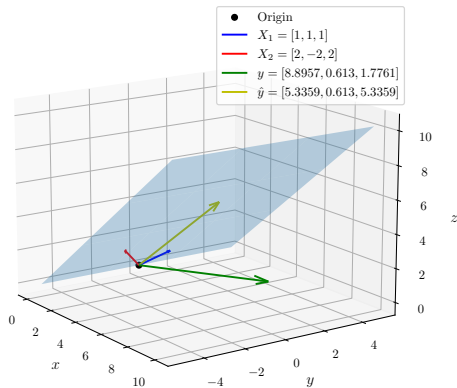


# Geometric Interpretation

Span of  $\left( \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix} \right)$

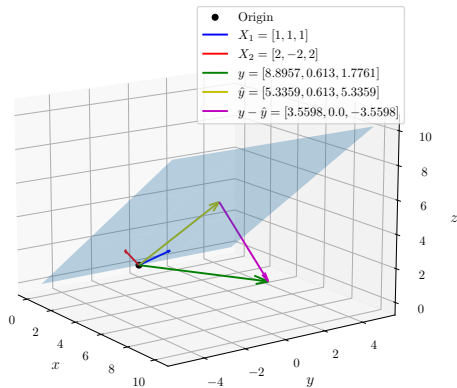


# Geometric Interpretation



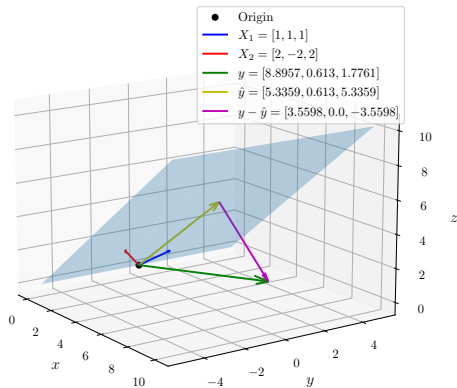
- We seek a  $\hat{\mathbf{y}}$  in the span of the columns of  $\mathbf{X}$  such that it is closest to  $\mathbf{y}$

# Geometric Interpretation



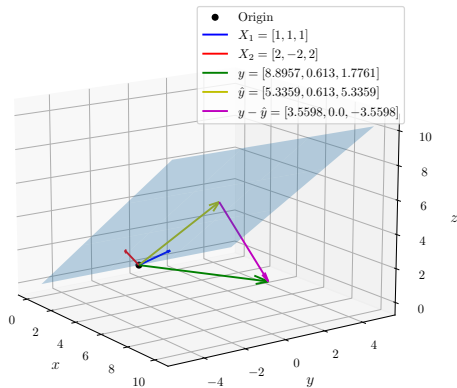
- This happens when  $\mathbf{y} - \hat{\mathbf{y}} \perp \mathbf{x}_j \forall j$  or  $\mathbf{x}_j^\top (\mathbf{y} - \hat{\mathbf{y}}) = 0$

# Geometric Interpretation



- This happens when  $\mathbf{y} - \hat{\mathbf{y}} \perp \mathbf{x}_j \forall j$  or  $\mathbf{x}_j^\top (\mathbf{y} - \hat{\mathbf{y}}) = 0$
- $\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\theta) = 0$

# Geometric Interpretation



- This happens when  $\mathbf{y} - \hat{\mathbf{y}} \perp \mathbf{x}_j \forall j$  or  $\mathbf{x}_j^\top (\mathbf{y} - \hat{\mathbf{y}}) = 0$
- $\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\theta) = 0$
- $\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\theta$  or  $\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$

# Dummy Variables and Multicollinearity

---

There can be situations where inverse of  $X^T X$  is not computable.

There can be situations where inverse of  $X^T X$  is not computable.  
This condition arises when the  $|X^T X| = 0$ .

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix} \quad (1)$$



## Multi-collinearity

There can be situations where inverse of  $X^T X$  is not computable.  
This condition arises when the  $|X^T X| = 0$ .

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix} \quad (1)$$

The matrix  $X$  is not full rank.

# Multi-collinearity

It arises when one or more predictor variables/features in  $X$  can be expressed as a linear combination of others

How to tackle it?

- Regularize

# Multi-collinearity

It arises when one or more predictor variables/features in  $X$  can be expressed as a linear combination of others

How to tackle it?

- Regularize
- Drop variables

# Multi-collinearity

It arises when one or more predictor variables/features in  $X$  can be expressed as a linear combination of others

How to tackle it?

- Regularize
- Drop variables
- Avoid dummy variable trap

## Dummy variables

Say Pollution in Delhi =  $P$

## Dummy variables

Say Pollution in Delhi =  $P$

$$P = \theta_0 + \theta_1 * \#Vehicles + \theta_2 * Wind\ speed + \theta_3 * Wind\ Direction$$

## Dummy variables

Say Pollution in Delhi =  $P$

$$P = \theta_0 + \theta_1 * \#Vehicles + \theta_2 * Wind\ speed + \theta_3 * Wind\ Direction$$

But, wind direction is a categorical variable.

# Dummy variables

Say Pollution in Delhi =  $P$

$$P = \theta_0 + \theta_1 * \#Vehicles + \theta_1 * Wind\ speed + \theta_3 * Wind\ Direction$$

But, wind direction is a categorical variable.

It is denoted as follows {N:0, E:1, W:2, S:3 }



# Dummy variables

Say Pollution in Delhi =  $P$

$$P = \theta_0 + \theta_1 * \#Vehicles + \theta_1 * Wind\ speed + \theta_3 * Wind\ Direction$$

But, wind direction is a categorical variable.

It is denoted as follows {N:0, E:1, W:2, S:3 }

Can we use the direct encoding?

# Dummy variables

Say Pollution in Delhi =  $P$

$$P = \theta_0 + \theta_1 * \#Vehicles + \theta_2 * Wind\ speed + \theta_3 * Wind\ Direction$$

But, wind direction is a categorical variable.

It is denoted as follows {N:0, E:1, W:2, S:3 }

Can we use the direct encoding?

Then this implies that  $S > W > E > N$

# Dummy Variables

N-1 Variable encoding

	Is it N?	Is it E?	Is it W?
N	1	0	0
E	0	1	0
W	0	0	1
S	0	0	0

# Dummy Variables

N Variable encoding

	Is it N?	Is it E?	Is it W?	Is it S?
N	1	0	0	0
E	0	1	0	0
W	0	0	1	0
S	0	0	0	1

Which is better  $N$  variable encoding or  $N-1$  variable encoding?

Which is better N variable encoding or N-1 variable encoding?

The N-1 variable encoding is better because the N variable encoding can cause multi-collinearity.

Which is better N variable encoding or N-1 variable encoding?

The N-1 variable encoding is better because the N variable encoding can cause multi-collinearity.

Is it  $S = 1 - (\text{Is it N} + \text{Is it W} + \text{Is it E})$

# Binary Encoding

N	00
E	01
W	10
S	11



# Binary Encoding

N	00
E	01
W	10
S	11

W and S are related by one bit.

## Binary Encoding

N	00
E	01
W	10
S	11

W and S are related by one bit.

This introduces dependencies between them, and this can cause confusion in classifiers.

## Interpreting Dummy variables

Gender	height
F	...
F	...
F	...
M	...
M	...

## Interpreting Dummy variables

Gender	height
F	...
F	...
F	...
M	...
M	...

Encoding

# Interpreting Dummy variables

Gender	height
F	...
F	...
F	...
M	...
M	...

Encoding

Is Female	height
1	...
1	...
1	...
0	...
0	...

# Interpreting Dummy Variables

## Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

## Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$$height_i = \theta_0 + \theta_1 * (\text{Is Female}) + \epsilon_i$$



## Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$$height_i = \theta_0 + \theta_1 * (\text{Is Female}) + \epsilon_i$$

We get  $\theta_0 = 5.9$  and  $\theta_1 = -0.7$

## Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$$height_i = \theta_0 + \theta_1 * (\text{Is Female}) + \epsilon_i$$

We get  $\theta_0 = 5.9$  and  $\theta_1 = -0.7$

$\theta_0 = \text{Avg height of Male} = 5.9$

## Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$$height_i = \theta_0 + \theta_1 * (\text{Is Female}) + \epsilon_i$$

We get  $\theta_0 = 5.9$  and  $\theta_1 = -0.7$

$\theta_0 = \text{Avg height of Male} = 5.9$

$\theta_0 + \theta_1$  is chosen based (equal to) on 5, 5.2, 5.4 (for three records).

## Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$$height_i = \theta_0 + \theta_1 * (\text{Is Female}) + \epsilon_i$$

We get  $\theta_0 = 5.9$  and  $\theta_1 = -0.7$

$\theta_0 = \text{Avg height of Male} = 5.9$

$\theta_0 + \theta_1$  is chosen based (equal to) on 5, 5.2, 5.4 (for three records).

$\theta_1$  is chosen based on 5-5.9, 5.2-5.9, 5.4-5.9

## Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

$$height_i = \theta_0 + \theta_1 * (\text{Is Female}) + \epsilon_i$$

We get  $\theta_0 = 5.9$  and  $\theta_1 = -0.7$

$\theta_0 = \text{Avg height of Male} = 5.9$

$\theta_0 + \theta_1$  is chosen based (equal to) on 5, 5.2, 5.4 (for three records).

$\theta_1$  is chosen based on  $5-5.9$ ,  $5.2-5.9$ ,  $5.4-5.9$   $\theta_1 = \text{Avg. female height } (5+5.2+5.4)/3 - \text{Avg. male height}(5.9)$

## Interpreting Dummy Variables

Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable

## Interpreting Dummy Variables

Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable

$$x_i = \begin{cases} 1 & \text{if } i \text{ th person is female} \\ -1 & \text{if } i \text{ th person is male} \end{cases}$$

## Interpreting Dummy Variables

Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable

$$x_i = \begin{cases} 1 & \text{if } i \text{ th person is female} \\ -1 & \text{if } i \text{ th person is male} \end{cases}$$

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i = \begin{cases} \theta_0 + \theta_1 + \epsilon_i & \text{if } i \text{ th person is female} \\ \theta_0 - \theta_1 + \epsilon_i & \text{if } i \text{ th person is male.} \end{cases}$$



## Interpreting Dummy Variables

Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable

$$x_i = \begin{cases} 1 & \text{if } i \text{ th person is female} \\ -1 & \text{if } i \text{ th person is male} \end{cases}$$

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i = \begin{cases} \theta_0 + \theta_1 + \epsilon_i & \text{if } i \text{ th person is female} \\ \theta_0 - \theta_1 + \epsilon_i & \text{if } i \text{ th person is male.} \end{cases}$$

Now,  $\theta_0$  can be interpreted as average person height.  $\theta_1$  as the amount that female height is above average and male height is below average.